第 37 卷第 2 期
录用定稿
计算机应用研究
Application Research of Computers
Vol. 37 No. 2
Accepted Paper

基于网络社团划分方法的多维数据聚类研究 *

吴行斌[1]，郭　强[1]，张林兵[1]，梁耀洲[1]，刘建国[2†]

(1. 上海理工大学 复杂系统科学研究中心，上海 200093; 2. 上海财经大学金融科技研究院，上海 200433)

摘　要：为了解决传统聚类方法在多维数据集中聚类效果不佳的问题，提出了将网络社团划分的方法，应用到多维数据聚类分析中。对于一个多维数据集，首先对分析对象进行特征提取，构建出每个对象的特征向量，通过计算皮尔森相关系数来度量不同特征向量之间的相似性，从而构建出一个相似性网络，采用 Blondel 算法对该网络进行社团划分达到聚类的效果。实验结果表明该方法可以在多维数据聚类中得到较好的聚类结果，准确率达到 92.5%，优于 k-means 算法的 75%。

关键词：聚类；多维数据；相似性；社团划分

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.07.0522

Multidimensional data clustering based on network community detection method

Wu Xingbin[1], Guo Qiang[1], Zhang Linbing[1], Liang Yaozhou[1], Liu Jianguo[2†]

(1. *Research Center for Complex Systems Science*, *University of Shanghai for Science & Technology*, *Shanghai* 200093, *China*; 2. *Institute of Fintech*, *Shanghai University of Finance & Economics*, *Shanghai* 200433, *China*)

Abstract: Since the traditional clustering method has an impoverished clustering effect on the multidimensional data, this paper used the clustering method based on community detection of complex networks to achieve better results. Firstly, the method extracted the features of original data and formed the feature vectors of each object for a multidimensional data set. Then it measured the similarity between different feature vectors by calculating Pearson correlation coefficients and constructed a similarity network. Finally, it used the Blondel algorithm which detects the community of the network to achieve the clustering effect. Experimental results show that this method can get better clustering results in multidimensional data clustering with an accuracy rate 92.5%, which is better than k-means algorithm with the accuracy rate 75%.

Key words: clustering; multidimensional data; similarity; community detection

0　引言

如何对多维数据进行分析，挖掘到数据背后所隐含的信息，这一问题日益引起了人们的广泛关注。有效而准确地聚类对于识别相似对象并挖掘不同类别之间异同具有重大意义。受维度灾难[1]的影响，传统的聚类方法在低维度数据空间上表现良好，而这些方法在多维度数据空间上往往不能得到好的聚类结果。寻找在多维数据中有效的聚类方法已成为聚类方法研究的重要方向。

聚类[2]作为统计分析的常用技术和数据挖掘的主要任务，一直被国内外学者广泛研究。Celebi 等人[3]针对 K-means 算法聚类中心初始位置太过敏感的问题，研究了已有解决该问题的初始化方法。Kriegel 等人[4]研究了基于密度的聚类，将聚类定义为比数据集的其余部分密度更高的区域，而在稀疏区域的对象通常被称为噪声和边界点。Tran 等人[5]提出了一种改进的 Dbscan 算法，解决了常规算法在检测相邻簇的边界对象时变得不稳定的问题。Ding 等人[6]针对类别之间的差异会掩盖子类间的差异的问题，提出了一种全新迭代聚类的框架，该框架可以在识别出大类之后，近一步识别出细微的差异，提供全面的聚类轨迹。Zhang[7]提出了一种新的勾股式模糊聚类方法，该方法能够解决事先没有精确给出准则权重的聚类问题。

除了寻找新的聚类算法，是否可以借鉴不同领域的方法应用到多维数据的分析中达到较好的聚类结果引起笔者深入的思考。经过数十年的发展，复杂网络的研究在社会科学[8,9]、计算机科学[10]、和生物科学[11]等众多领域中有了广泛应用。其中 2002 年 Girvan 和 Newman 指出复杂网络中普遍存在聚类特性[12]，并于 2004 年提出了一种发现这些社团的 GN 算法[13]。在此之后，大量学者对复杂网络中的社团划分问题进行了大量研究，Papadopoulos 等人首次在社会媒体的语境中构造了社团的概念[14]。Xie 等人[15]总结了最先进的重叠社团划分算法质量度量的基准，发现在不同重叠密度的网络中，不同算法具有不同的稳定性。社团划分至今仍存在一些不明确的问题，如社团本身的定义未统一，算法的验证及其性能的比较没有通用的标准。Fortunato 等人[16]针对这些问题，提出了一套指导流程，指出了现有流行算法的优点和缺陷，并给出了不同算法的使用方向。杨等人[17]基于拉普拉斯矩阵多重特征向量提出了一种有向网络社团结构划分算法，同时研究了动态网络中新社团成员的演化特性[18]。

网络中的的社团与常规聚类算法中的类的概念是一致

收稿日期：2018-07-13；**修回日期：**2018-08-30　**基金项目：**国家自然科学基金资助项目（61773248，71771152）

作者简介：吴行斌（1994-），男，安徽黄山人，硕士研究生，主要研究方向为复杂网络、数据挖掘；郭强（1975-），女，教授，博士，主要研究方向为复杂网络、数据挖掘、科学知识图谱分析；张林兵（1994-），男，硕士研究生，主要研究方向为复杂网络、数据挖掘；梁耀洲（1994-），男，硕士研究生，主要研究方向为复杂网络、数据挖掘；刘建国（1979-），男（通信作者），教授，博士（后），主要研究方向为网络科学、商务智能、在线用户行为分析（liujg004@ustc.edu.cn）.

的，因此将社团划分算法应用到聚类分析中具有可行性。本文关注传统聚类方法在多维数据上聚类效果不佳的问题，结合复杂网络社团划分的思想，将网络划分的方法应用到多维数据的聚类中。结果发现相比传统的聚类方法，网络的划分方法得到的聚类结果准确率更高，分类结果更准确。

1　相关理论

针对多维数据集进行分析，本文设计了一套处理流程，首先对原始数据集进行预处理，去除异常值、空缺值等异常数据，根据分析需求和实际场景对研究对象进行特征的获取，随后计算不同研究对象的特征之间的相关系数从而构造出相似性网络，再对该网络运用 Blondel 算法进行社团划分，达到聚类效果，最后可对聚类的结果进行进一步分析。实验过程如图 1 所示。

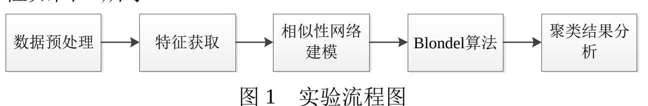


图 1　实验流程图

Fig. 1　Experimental flow graph

1.1　网络的构建

网络的节点代表聚类的对象，边代表对象之间的相关性。任意两个对象 i 和对象 j 的皮尔森相关系数由式(1)定义。

$$corr(D_i,D_j)=\frac{\mathrm{cov}(D_i,D_j)}{\sqrt{\mathrm{cov}(D_i,D_i)\mathrm{cov}(D_j,D_j)}} \quad (1)$$

如果皮尔森相关系数 $|corr(D_i,D_j)|\geq\theta(\theta\in[-1,1])$，就认为节点 i 和 j 之间有连边，这里的 θ 即阈值点。选取合适的阈值可以构造出具有明显拓扑结构的网络，有利于后续研究展开。

1.2　Blondel 社团划分算法

Newman 等人在 2004 年首先提出了模块度的概念，模块度是一种衡量社团划分质量的指标。如式(2)所示。

$$Q=\sum_{ij}\left[\frac{A_{ij}}{2m}-\frac{k_ik_j}{(2m)*(2m)}\right]\delta(c_i,c_j) \quad (2)$$

对于所研究的网络而言，式中 A_{ij} 表示网络的邻接矩阵，k_i 表示节点 i 的度，m 是网络的边数，　C_i 是节点 i 所属的社团，当 $C_i=C_j$ 时，$\delta(c_i,c_j)=1$，否则为 0；Q 值在 0 和 1 之间。Q 值越大说明社团划分出的结构越有效。

Blondel 等人[19]在 2008 年提出一种基于模块度的快速聚类算法。它主要目标是不断划分社团使得划分后的整个网络的模块度不断增大，划分后的网络模块度越大，说明社团划分的效果越好。文献中提出当节点 i 被划分到社团 C 中去时，社团 C 的模块度增益计算公式如式(3)所示。

$$\Delta Q=\left[\frac{\sum_{in}+2k_{i,in}}{2m}-\left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right]-\left[\frac{\sum_{in}}{2m}-\left(\frac{\sum_{tot}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right] \quad (3)$$

其中 $\sum_{in}$ 是 C 社团中的内部连接权重的总和，$\sum_{tot}$ 是 C 社团指向节点的所有连接权重的总和，$k_{i,in}$ 是节点 i 与社团中其他节点连接权重的总和。

算法描述如下：

a)将网络中每个节点视作单独的一个社团；

b)对每一个节点，将每个节点尝试划分到与其邻接的点所在的社团中，计算此时的模块度，判断划分前后的模块度的差值 ΔQ 是否为正数，若为正数，则接受此次划分，若不为正数，则放弃本次划分；

c)重复步骤 b)，直到社团划分后的模块度不再增大为止；

d）将步骤 c)得到的社团结构中的每个社团视为新的节点，构造新的网络，重复执行步骤 b)c)，直到社团的结构不再改变为止。

2　实验及分析

2.1　实验数据

本文采取了国内某条公交线路 2017 年 9 月 133 名司机、55 辆车的数据，针对该线路的公交司机进行聚类分析研究。数据集包含有记录日期，记录时间，车辆 ID，驾驶员 ID，位置经度，位置纬度，CAN 速度，等 17 个字段。

2.2　数据清洗

首先对原始数据进行清洗。针对实际数据在收集的过程中出现的缺失数据，本文根据实际业务分析，采用了 hot deck[20]填补法，即就近补齐法。

本文对原始数据进行了逻辑错误检测。例如在行车路程轨迹段，有部分车辆的车速数据超过正常范围达到了 120km/h，针对该非常规数据的处理方法是删除这个轨迹段。对于一些数据缺失较多的字段，例如司机制动踏板开度字段信息缺失率接近 100%，本文进行舍弃处理。对于一些数据字段中存在异常值的情况，例如电机转速达到 16 383 转速的异常值，则直接将异常值替换为空白值，将空白值就近补齐。

2.3　特征获取

为了让司机的行为特征具有可比性，本文设定一种具体的分析场景，即司机驾驶公交车从起点到终点为完整的一趟行车记录，以此作为每个司机的行为特征计算的基准。

首先将清洗后的数据按照趟的场景进行切分，在切片后的数据中提取出司机每一趟驾驶公交车的行为特征。根据有效数据和场景推断，本文选取的特征为加速度绝对值平均值；加速度绝对值标准差；车速平均值；车速标准差；电子刹车使用概率；脚刹使用概率；油门踏板百分比平均值；油门踏板百分比标准差；车速中位数共九个特征。其中对于车辆加速和减速造成加速度正负值问题，本文对加速度取绝对值进行计算。

2.4　相似性网络建模

司机驾驶公交车行驶一趟公交线路的属性可以表示为一个 n 维向量 $D_i=\{x_{i1},x_{i2},\dots,x_{in}\}$，其中 i 代表司机行驶的第 i 趟，$x_{i1},x_{i2},\dots,x_{in}$ 表示第 i 趟行驶过程中有关司机驾驶的 n 个行为特征属性。依据不同趟的行为特征的相似性进行网络构建，本文采取皮尔森相关系数来度量不同司机之间的特征相似性。由式(1)得

$$corr(D_i,D_j)=\frac{\mathrm{cov}(D_i,D_j)}{\sqrt{\mathrm{cov}(D_i,D_i)\mathrm{cov}(D_j,D_j)}} \quad (1)$$

其中：D_i 和 D_j 分别为第 i 趟和第 j 趟驾驶行为特征属性向量，若两趟行驶过程中驾驶行为特征在各分量特征上具有一定的相似性，则给这两趟驾驶行为特征向量构建一条连边。根据式(1)计算相似性，设定阈值 0.15，当不同趟的驾驶行为特征向量间的相似性大于该阈值时，建立连边。最终得到的网络包含 879 个节点，386760 条连边。节点平均度为 440.

2.5　聚类结果及准确性度量

根据皮尔森相关系数构建出的网络，采取了 Blondel 算法进行社团划分，得到社团数量为 3 类，其中第 1 类包含 679 趟驾驶行为，101 个不同司机，第 2 类包含 199 趟驾驶行为，40 个不同司机，第 3 类包含 1 趟驾驶行为，其中只包含 1 个司机。

在每一个类别中，以每趟行车记录的行为特征为聚类对象，即对一个司机而言，若他的行为特征具有稳定性，他的

每趟行车记录都会聚到同一个类别中，若存在司机行为发生变化的情况下，则会导致行车记录聚到不同的类中。因此本文定义了一种司机平均分类准确性指标，如果每趟行车记录司机行为都能有效分到同样的类中，表明分类准确性最高。平均分类准确性的定义公式如式(4)所示。

$$p_c = \frac{1}{m}\sum_{i=1}^{m}\frac{\max\{n_i^{c_l}\}}{n_i} \quad C_l = 1,2,3 \tag{4}$$

其中：n_i 为司机 i 行驶的总趟数，C_l 为分类结果，$n_i^{c_l}$ 为第 C_l 类中该司机的行车趟数，$\max\{n_i^{c_l}\}$ 表示该司机行驶记录最多的趟数聚到某类中的数量。理想状态下，单个司机的驾驶记录都会分到一个类中，那么 $\max\{n_i^{c_l}\}/n_i = 1$。$m$ 为司机总数。最终对所有司机的分类准确性计算均值，得到平均分类准确性。司机聚类准确性示例如表 1 所示。

表 1　司机聚类准确性示例表

Table 1　The sample of driver clustering accuracy

司机	车	车速平均值	加速度绝对值平均值	油门踏板百分比平均值	聚类结果
司机 1	车 1^1	21	0.22	17	第一类
	车 1^1	19	0.19	14	第一类
	车 1^1	18	0.21	16	第一类
司机 2	车 2^2	20	0.24	8	第二类
	车 2^2	19	0.22	9	第二类
	车 2^2	18	0.21	8	第二类
	车 3^2	21	0.2	7	第二类
司机 3	车 4^3	19	0.21	17	第一类
	车 4^3	22	0.25	19	第一类
	车 4^3	18	0.23	16	第一类
	车 5^3	19	0.22	8	第二类
	车 5^3	20	0.23	7	第二类

以表 1 为例，司机 1 仅驾驶了车 1，这三趟驾驶行为差异较小，均被分到类 1 中，那么该司机的分类准确性为 100%。司机 2 驾驶过车 2 和车 3，驾驶两辆车行为差异不大，司机 2 的所有记录都分到了类 2，分类准确性为 100%。司机 3 驾驶过车 4 和车 5，驾驶车 4 的三趟记录分到了类 1，驾驶车 5 的两趟记录分到了类 2，分类准确性为 3/5*100%=60%。最终得到的 3 个司机的平均分类准确性为（100%+100%+60%）/3=86.67%。

为了对比上述方法的优缺点，本文设置聚类簇的个数为 3，采用 K-means 算法对划分好的“趟”数据进行聚类分析，得到 K-means 算法的结果。根据本文的分类准确性验证方法，验算 K-means 算法计算的平均准确性为 75%，而 Blondel 算法在阈值确定的情况下准确率为 92.5%。相比 K-means 算法准确率明显提高。

3　结束语

本文提出将网络社团划分的算法——Blondel 算法应用到多维数据的聚类分析中。通过对实际公交运营数据的分析，设计出具体的公交车辆从起点到终点为一趟的分析场景。根据每趟驾驶行为的属性向量之间的相似性构造出网络。采用 Blondel 算法对该网络进行社团划分，得到聚类的结果。通过提出的准确率度量方法进行准确率度量，结果显示网络社团划分算法准确率达到 92.5%，优于 K-means 聚类算法 75%的准确率。

多维数据的聚类分析在众多领域有着广泛的应用前景。本文使用网络的社团划分方法对多维数据进行聚类，但是对于更高维度的数据上的聚类效果还需继续进一步研究。同时对于多维数据上的聚类算法可以从更多的角度思考与研究。

参考文献：

[1] Zimek A, Schubert E, Kriegel H P. A survey on unsupervised outlier detection in high‐dimensional numerical data [J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2012, 5 (5): 363-387.

[2] Sarstedt M, Mooi E. A concise guide to market research [M]. 2nd ed. Berlin：Springer, 2014: 273-324.

[3] Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods for the k-means clustering algorithm [J]. Expert Systems with Applications, 2013, 40(1): 200-210.

[4] Kriegel H P, Kröger P, Sander J, *et al.* Density-based clustering [J]. Data Mining & Knowledge Discovery, 2011, 1(3): 231-240.

[5] Tran T N, Drab K, Daszykowski M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters [J]. Chemometrics and Intelligent Laboratory Systems, 2013, 120(15): 92-96.

[6] Ding Hongxu, Wang Wanxin, Califano A. IterClust: a statistical framework for iterative clustering analysis [J]. Bioinformatics, 2018, 34 (16): 2865-2866.

[7] Zhang Xiaolu. Pythagorean fuzzy clustering analysis: a hierarchical clustering algorithm with the ratio index-based ranking methods [J]. International Journal of Intelligent Systems, 2018, 33(9): 1798-1822.

[8] 任卓明, 刘建国, 邵凤, 等. 复杂网络中最小 k-核节点的传播能力分析 [J]. 物理学报, 2013, 62 (10): 108902. (Ren Zhuoming, Liu Jianguo, Shao Fng, *et al.* Analysis of the spreading influence of the nodes with inimum k-shell value in complex networks [J]. Acta Physica Sinica, 2013, 62 (10): 108902)

[9] 杨剑楠，刘建国，郭强. 基于层间相似性的时序网络节点重要性研究 [J]. 物理学报，2018, 67(4): 48901-048901. (Yang Jiannan, Liu Jianguo, Guo Qiang. Node importance idenfication for temporal network based on inter-layer similarity [J]. Acta Physica Sinica, 2018, 67 (4): 48901-048901.)

[10] 胡小军，郭强，杨凯，等. 基于相对熵的多属性作者学术影响力排名研究 [J]. 电子科技大学学报，2018, 47(2): 279-285. (Hu Xiaojun, Guo Qiang, Yang Kai, *et al.* Multi-attribute researcher academic influence ranking based on relative entropy [J]. Journal of University of Electronic Science and Technology of China, 2018, 47(2): 279-285.)

[11] Bloomfield N J, Knerr N, Encinas V F. A comparison of network and clustering methods to detect biogeographical regions [J]. Ecography, 2018, 41(1): 1-10.

[12] Girvan M, Newman M E. Community structure in social and biological networks [J]. Proceedings of The National Academy of Sciences, 2002, 99 (12): 7821-7826.

[13] Newman M E, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69(2): 026113.

[14] Papadopoulos S, Kompatsiaris Y, Vakali A, *et al.* Community detection in social media [J]. Data Mining and Knowledge Discovery, 2012, 24 (3): 515-554.

[15] Xie Jierui, Kelley S, Szymanski B K. Overlapping community detection in networks: the state-of-the-art and comparative study [J]. ACM Computing Surveys, 2013, 45 (4): 43.

[16] Fortunato S, Hric D. Community detection in networks: a user guide [J]. Physics Reports, 2016, 659 (11): 1-44.

[17] 杨凯，郭强，刘晓露,等. 基于多重特征向量的有向网络社团结构划分算法 [J]. 电子科技大学学报，2016, 45(6): 1014-1019. (Yang Kai, Guo Qiang, Liu Xiaolu, *et al.* Detecting community structure in directed

networks via multiple eigenvectors [J]. Journal of University of Electronic Science and Technology of China, 2016, 45(6): 1014-1019, 1032.)

[18] Yang Kai, Guo Qiang, Li Shengnan, *et al.* Evolution properties of the community members for dynamic networks [J]. Physics Letters A, 2017, 381(11): 970-975.

[19] Blondel V D, Guillaume J L, Lambiotte R, *et al.* Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008

[20] Myers T A. Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data [J]. Communication Methods & Measures, 2011, 5(4): 297-310.